ICSI /ThisL status report

December 1997

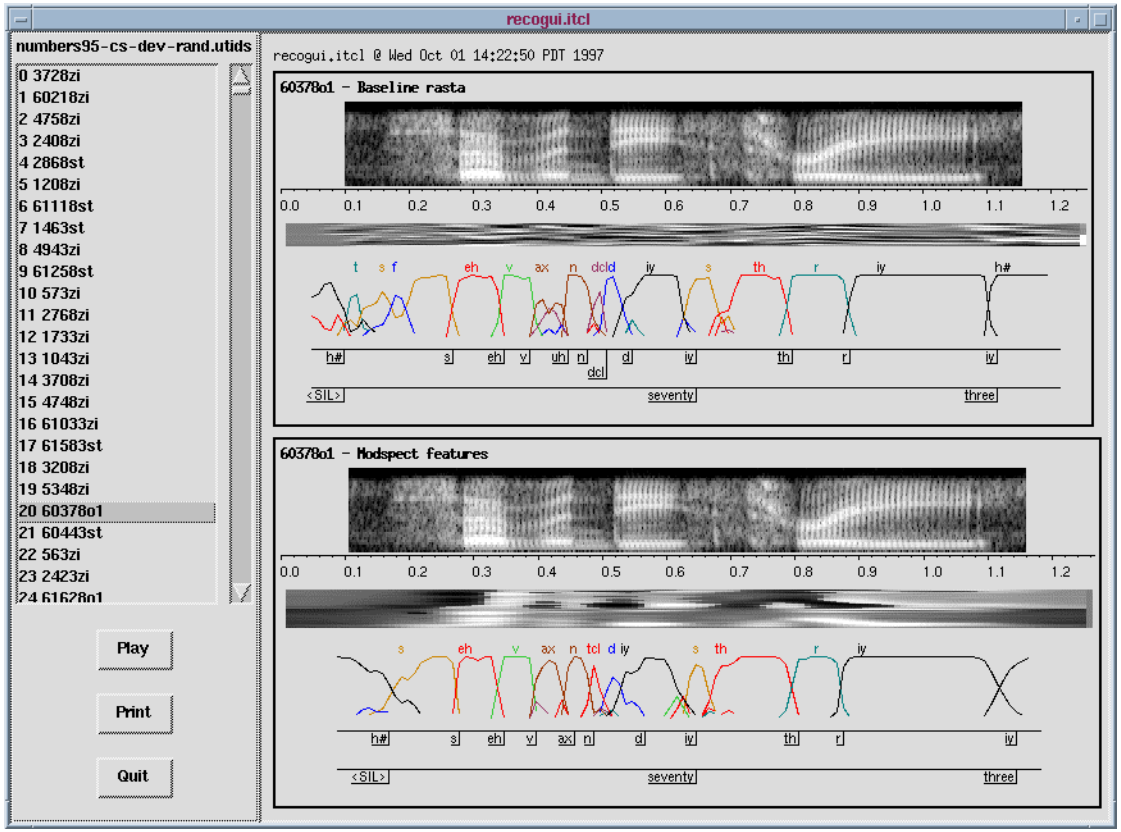
Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

1. **Software tools & packages**
2. **Speech/nonspeech separation**
3. **Speech in reverberation**

1 Software tools & packages: ICSI Speech Recognition system

- **New components:**
  - **`feacalc`**: enhanced RASTA (I/O formats, options)
  - **`pfile_utils`**: comprehensive manipulations (editing, stats, etc.)
- **Portable package:**
  - first test: bring up recognizer at IDIAP
- **Visualization**
  - [incr Tcl] classes for display
  - recognizer visualization...

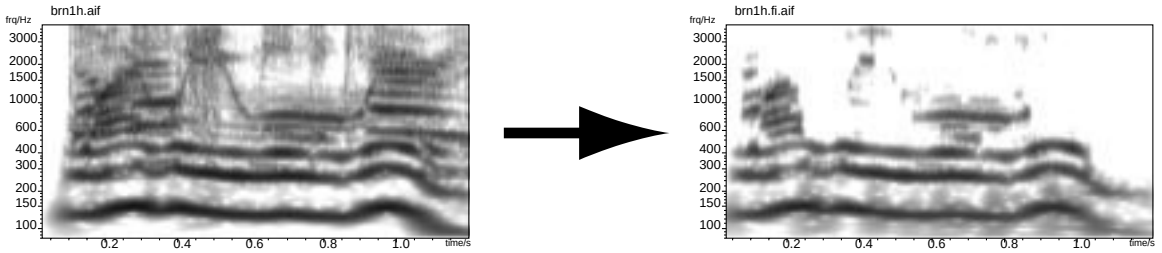
Software tools & packages: `recogui` visualization



- modular objects for reuse
- simple configuration files
- broad use within ICSI

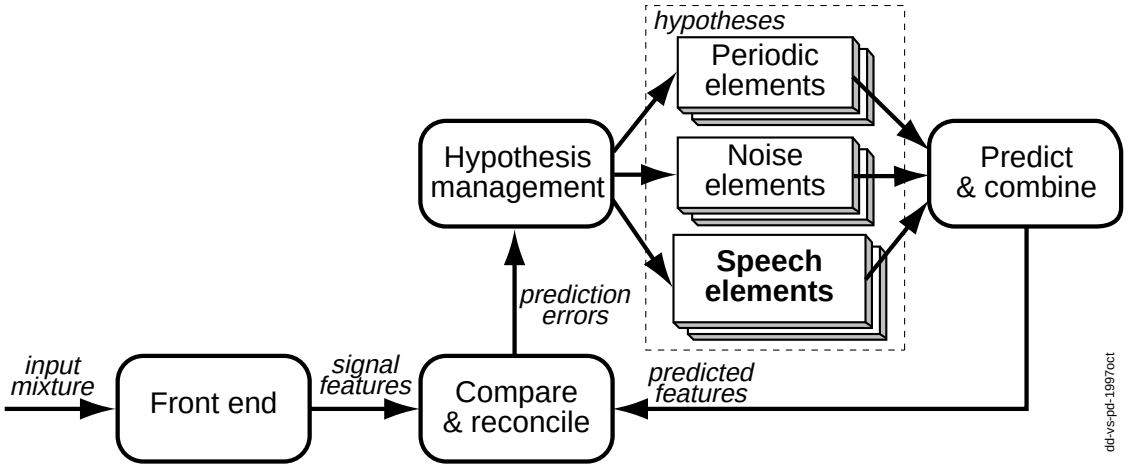
2 Speech vs. nonspeech: Comp. Aud. Scene Analysis for ASR

- **For handling sound mixtures, attempt to estimate individual sound sources**
  - listeners do this transparently
- **Previous approach (Weintraub...)**
  - ‘enhance-then-recognize’:
    extract by periodicity, resynthesize, recognize
- **But...**
  - problems with ‘holes’
  - which cues to separate speech?
    ...doesn’t exploit knowledge of speech structure



Prediction-driven CASA and ASR

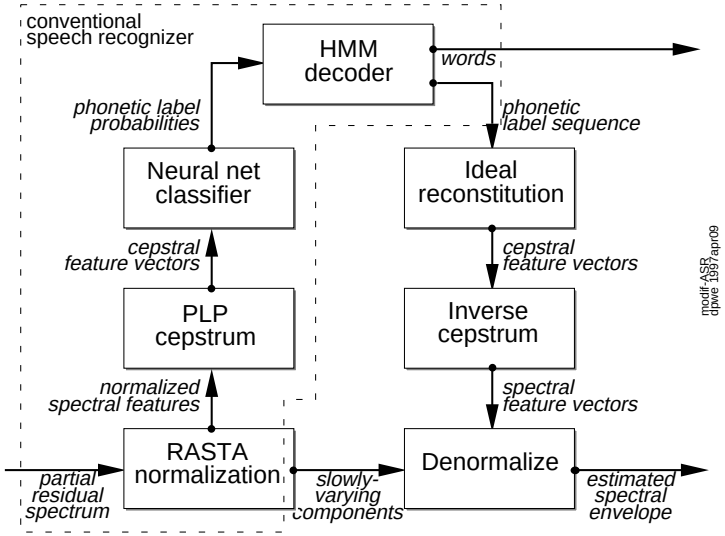
- **Prediction-driven CASA:**
  - don't *derive*, but *construct* an explanation consistent with observations



  - need to express 'predictions' in signal domain
  - iterate over each component

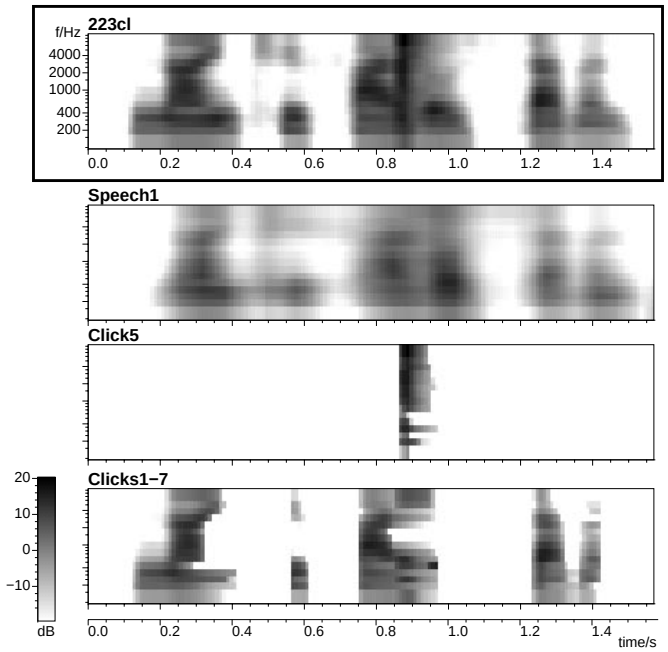
- **Components make *projections***
  - into e.g. 'the space of all speech sounds'

A Speech Hypothesis module



- **Want to exploit constraints of decoder**
- **Invert each stage of speech recognizer**
  - classifier by? trained estimator
  - normalization by: recovering from input

Preliminary results



- **Prediction shortfall dominates result**
  - improve inverse classification
  - more normalization
- **To complete iteration:**
  - need p(q|X,M)
- **Initial separation by f0?**

3 Speech-in-reverberation

- **Modest reverb has severe impact**
  (RT = 0.5s, D/R ≈ 0 dB)
- **Information/combination at various timescales**
  - modulation spectral features, syllable units
  - combine results at utterance level (Nbests)
  - combine results at syllable level
    (HMM decomposition, [Dupont & Bourlard '97]

WER%	Clean speech	Reverb (6 dB SNR)
Baseline (Rasta-PLP8)	6.8	27.8
ModSpec Syllable base	9.8	30.9
Utterance-level combin'n	5.5	19.6
Syllable-level combin'n	5.4	18.6

- **2 pass decoder to avoid state explosion**
  - lattice output for compatibility